Mathematical Statistics
Stockholm University

Retrospective Ancestral Recombination Graphs with Applications to Gene Mapping

Linda Hartman and Ola Hössjer

Research Report 2007:7

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2007:7**,
http://www.math.su.se/matstat

Retrospective Ancestral Recombination Graphs with Applications to Gene Mapping

Linda Hartman and Ola Hössjer

20070307

Abstract

Rare diseases are often investigated in case-control studies. For the purpose of gene mapping, sampled case and control genotypes are compared at a set of marker locations. Closely linked marker loci can be handled by modeling the genealogy of the sample. We present such a model, which splits the chromosomes into subpopulations. In this way the model accounts for the ascertainment process, where cases are typically over sampled. The model is used for multipoint gene mapping by means of a LOD score. The LOD score copes with arbitrary phenotypes and genetic models, allows for neutral mutations, and adapts to marker allele frequencies. Under certain model approximations we develop a permutation based test that is computationally feasible, even when haplotype phase is unknown.

KEY WORDS: Association analysis; multipoint; unknown haplotype phase; ascertainment; case-control study; coalescent; identical by descent (IBD); sampling; SNP; LOD score;

1 Introduction

For the purpose of gene mapping on a fine scale, the use of population-based association studies is popular. Due to the higher number of meioses, with possible recombination, between the most recent common ancestor (MRCA) and today's apparently unrelated individuals, these studies yield higher resolution than is possible in family-based linkage studies.
Although early association studies tested for association between disease and each marker separately, effort is nowadays put into finding efficient methods that evaluate linkage disequilibrium (LD) over more than one locus within a region. This article presents a LOD score for population based multipoint association studies, which takes a retrospective sampling scheme into account. The basis of the LOD score is to use a model for chromosome genealogies that handles non-random ascertainment.
Early multipoint approaches to gene-mapping, such as Terwilliger (1995), combined information from many markers, but did not include dependence across markers in the analysis. Genuine multipoint likelihood methods such as McPeek and Strahs (1999) and Service et al. (1999) condition on (possibly unknown) ancestral haplotypes to calculate the probabilities that today's haplotypes are identical by descent (IBD) with the variant founder at the disease locus.
In a region in LD, alleles at different loci are dependent. The genealogical history of the population sample determines that dependence, and could thus be used for mapping purposes. Griffiths and Marjoram (1997) constructed the Ancestral Recombination Graph (ARG) to model how a population of chromosome sequences are related to each other, through coalescence, recombination and mutations. The ARG extends the Wright-Fisher model for coalescence to allow also for recombinations, which are of immediate interest, as they break down LD.
Due to recombinations the genealogies differ between different chromosomal positions. For each position the ARG defines a marginal genealogical tree. This could be used for gene mapping, by searching for the chromosomal position where the marginal tree can discriminate cases from controls, i.e. where a majority of cases are on the same branch of the genealogical tree. Unfortunately however, the ARG is not known, and of the infinitely many ARGs compatible with genotype data, many have comparable likelihoods. Larribe et al. (2002) use importance sampling in an ARG model as a method to estimate disease locus, but the method is hampered by the computational demand. In practice, only ARGs fulfilling simple approximations are computationally tractable for mapping.
In the definition of the ARG by Griffiths and Marjoram (1997) (defined as

a process in time) or Wiuf and Hein (1999) (defined as a process along the chromosomal region), the same model for coalescence, recombinations and mutations are used for all chromosomes in the sample. However, in association studies, there is an underlying hypothesis that cases are descendents from one (or a few) founders carrying the mutation. The genealogy of the cases is then qualitatively different from that of the controls. In a retrospective study, the sample is typically not a random sample from the population. In general, the mutated allele is over-represented, although its exact proportion is in general unknown.

This article presents an extension of the ARG, which, by separating the chromosomes into two sub-populations of mutated and unmutated chromosomes respectively, models the evolutionary process of the two types of chromosomes differently. Zöllner and von Haeseler (2000), later elaborated by Wang and Rannala (2004, 2005), put forward a similar course of action, in introducing an ARG with subpopulations. The focus of these models is on simulation, and to examine the performance of single locus association tests for chromosomes simulated under different scenarios. No multilocus gene-mapping algorithm is developed.

Apart from catching the different genealogical behaviour of mutated and unmutated chromosomes, an ARG with substructure also opens up for approximations that could be differently tailored for the two subpopulations. Thus easier computations are facilitated, while still catching the important features of mutated and unmutated chromosomes respectively. Although in real studies the mutation status of the chromosomes is not known, the distribution of mutated chromosomes can easily be calculated conditional on disease status. For genetic models with full penetrance and imprinting, disease status determines whether a chromosome is mutated or not. Otherwise mutation status is treated as a hidden variable with known distribution conditional on disease status. The retrospective ARG is therefore a useful model under case-control sampling, where ascertainment is on disease status.

Our main result is to present how our retrospective ARG can be used to calculate a likelihood and LOD score. Thus we connect the retrospective ARG to a multipoint gene mapping algorithm. The basis of calculating the likelihood is to use the retrospective ARG to model which regions that chromosomes share identical by descent (IBD). The retrospective ARG is general, but for computational reasons we suggest model simplifications and present a special case for which the LOD score is computationally achievable. We further show how exact p-values can be estimated by a permutation procedure, which is computationally feasible for binary phenotypes. Just as in Terwilliger (1995) and Service et al. (1999), the basis of the simplifications is to assume star topology for cases. However, our setting is more general, as

the resulting retrospective likelihood handles more markers, arbitrary phenotypes and genetic models, allows for neutral mutations, and adapts to marker allele frequencies. Further, unknown haplotype phase is handled at almost no extra computational cost. This is a marked improvement compared to Terwilliger (1995) and Service et al. (1999) since phased marker haplotypes are seldomly observed. Instead, it is the unphased multilocus diplotypes that are observed, and except for rarely collected data sets, phase cannot be resolved unambiguously. Thus, in most haplotype based analyses, the haplotypes constitute a covariate not fully observed. Not handling the retrospective sampling scheme in the analysis, may then introduce bias, see Thomas et al. (2003).

2 A LOD Score for Association Studies

The purpose of gene mapping is to test if a certain (small) chromosome region harbours the disease locus τ and/or estimate τ. The region of interest is normalized as a unit interval $[0,1]$ in terms of genetic or physical map distance. The hypothesis testing problem of interest is

$$\begin{aligned} H_0: &\quad \tau \notin [0,1], \\ H_1: &\quad \tau \in [0,1]. \end{aligned}$$

To investigate this, a subset of m individuals with phenotypes $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ is sampled. For each individual DNA is registered at a number of markers with positions $0 \leq x_1 < \ldots < x_K \leq 1$. Let $h_{2v-1} = (h_{2v-1,k})_{k=1}^K$ and $h_{2v} = (h_{2v,k})_{k=1}^K$ be the two homologous haplotypes of Individual v and $\boldsymbol{h} = (h_i)_{i=1}^n$ the collection of all $n = 2m$ haplotypes. In general, because of phase uncertainty, (h_{2v-1}, h_{2v}) is not known for v but rather the unphased multilocus genotype g_v. Write $\boldsymbol{g} = (g_1, \ldots, g_m)$ for the collection of all unphased multilocus genotypes. Based on marker data $\boldsymbol{g}$ and phenotypes $\boldsymbol{Y}$ we compute a test statistic $Z(x)$ for the point wise test H_0 versus $H_1^x :' \tau = x'$ and reject H_0 when $Z(x)$ is large. Then

$$Z_{\max} = \max_{0 \leq x \leq 1} Z(x)$$

is a global test statistic for testing H_0 versus H_1, with large values of $Z_{\max}$ leading to rejection of H_0. Alternatively, we may estimate the disease locus as $\hat{\tau} = \arg\max_{0 \leq x \leq 1} Z(x)$ and compute an associated confidence region.

The test statistic $Z(x)$ should be large when diseased individuals, or individuals with quantitative phenotypes indicating disease, tend to share DNA around x more often than expected by chance. This is so since under H_1^x, the

mutated chromosome is segregated in close vicinity of x down to all mutated chromosomes of the sample.
To this end, we define the retrospective likelihood

$$L(x;\xi) = P_x(\boldsymbol{g}|\boldsymbol{Y}) \tag{1}$$

of genotype data given phenotypes, where P_x is probability calculated under H_1^x. By conditioning on $\boldsymbol{Y}$ we don't need to know the sampling mechanism, as long as it is a function of $\boldsymbol{Y}$ only. This is an advantage, since the sampling scheme is often unknown in practice, see e.g. Kraft and Thomas (2000). All nuisance parameters that involve recombination, mutation, population growth and penetrance of the disease are contained in ξ. Assuming ξ is known or put to an apriori reasonable value, as test statistic we use the LOD score

$$Z(x) = \log_{10} \mathrm{LR}(x) = \log_{10} \frac{L(x)}{L(\infty)}, \quad 0 \leq x \leq 1. \tag{2}$$

Here $L(\infty)$ denotes the retrospective likelihood under H_0, since then τ is regarded as unlinked to $[0,1]$, expressed formally as $\tau = \infty$. Hence $Z(x)$ is the tenth logarithm of the likelihood ratio $\mathrm{LR}(x)$ obtained when testing H_0 against H_1^x.

To assess the statistical significance of an observed maximal LOD score $Z_{\max} = z_{\max}$, we use permutation testing. Given any permutation γ of $\{1,\ldots,m\}$, let $Z_{\max,\gamma}$ be the maximal LOD score based on a retrospective likelihood $P_x(\boldsymbol{g}|\boldsymbol{Y}_\gamma)$, where $\boldsymbol{Y}_\gamma = (Y_{\gamma(1)},\ldots,Y_{\gamma(m)})$ is the phenotype vector permuted according to γ. The p-value based on Q randomly chosen permutations $\gamma_1,\ldots,\gamma_Q$ is then then $\alpha(z_{\max})$, where

$$\alpha(z) = \frac{1}{Q}\sum_{i=1}^{Q} I(Z_{\max,\gamma_i} \geq z) \tag{3}$$

and $I(D)$ is the indicator function of the event D.

3 A Retrospective Recombination Graph

The likelihood that we propose handles dependence between and along the chromosomes, by means of an ARG corrected for ascertainment. The ARG models the genealogy from today's generation back until the founder generation. That gives us the kinship relations of today's chromosomes, and thus a model for the dependencies. In order to compute the likelihood and LOD score, we will sum over the genealogies consistent with data when $x \in [0,1]$.

To this end, consider the evolvement of a diploid (human) population during G non-overlapping generations consisting of N_t haploids or $N_t/2$ individuals t generations ago, $t = 0, 1, \ldots, G$. The current size of the population is $N/2 = N_0/2$ and G is the founder generation. A disease causing mutation occurred $t = G_M$ generations ago on one chromosome at an unknown locus τ of the genome. The mutated material in close vicinity of τ has spread until present time so that N_{Mt} chromosomes have the mutated and disease causing allele B in Generation t, whereas the remaining $N_{Ut} = N_t - N_{Mt}$ unmutated chromosomes have the normal allele b. The current numbers of mutated and unmutated chromosomes are $N_M = N_{M0}$ and $N_U = N_{U0}$ respectively. For instance, for an exponentially growing population,

$$\begin{aligned} N_t &= N \exp(-\kappa t), \\ N_{Mt} &= N_M \exp(-\kappa_M t), \end{aligned} \tag{4}$$

where $\kappa > 0$ and $\kappa_M = \log(N_M)/G_M$ quantify the rate of exponential growth per generation.

We will assume that N_M and G_M (and hence also N and G) are large. Then the genealogy at x is conveniently approximated by a coalescence tree $\mathcal{T}(x)$ Kingman (1982a,b). The coalescence tree varies with x due to recombinations and the whole collection $\mathcal{A} = \{\mathcal{T}(x);\, 0 \leq x \leq 1\}$ is referred to as an ancestral recombination graph (ARG), see e.g. Hudson (1983) and Griffiths and Marjoram (1997).

Measuring time backwards and continuously in units of G_M generations, we let $N(t) = N_{G_M t}$, $N_M(t) = N_{M,G_M t}$ and $N_U(t) = N(t) - N_M(t)$, be the sizes of the total, mutated and unmutated populations of chromosomes on the new time scale, $0 \leq t \leq T = G/G_M$.

Figure 1 shows an ARG for $n = 13$ chromosomes. For a disease locus in $\tau = 0.36$, the lineages carrying the mutation are marked in dashed gray. Following Griffiths and Marjoram (1997), we define the ARG as a process in time rather than along the chromosome. We write it as a time homogeneous Markov process $\mathcal{A} = \{A(t); 0 \leq t \leq T\}$, where $A(t)$ describes the ancestry across $[0, 1]$ of the given sample of n chromosomes at time t. As t increases, $A(t)$ make jumps at discrete points in time due to coalescence or recombination events.
Each vertex corresponds to a coalescent or recombination event and each edge e is a line of descent between two such events. Whenever a recombination event occurs, the recombination point x is marked with an arrow to the vertex in the graph. The marginal coalescence tree $\mathcal{T}(x')$ at x' is obtained from $\mathcal{A}$ by following all n lineages at time 0 and, whenever a recombination vertex marked x is passed, take the left edge if $x' < x$ and the right edge if $x' \geq x$.

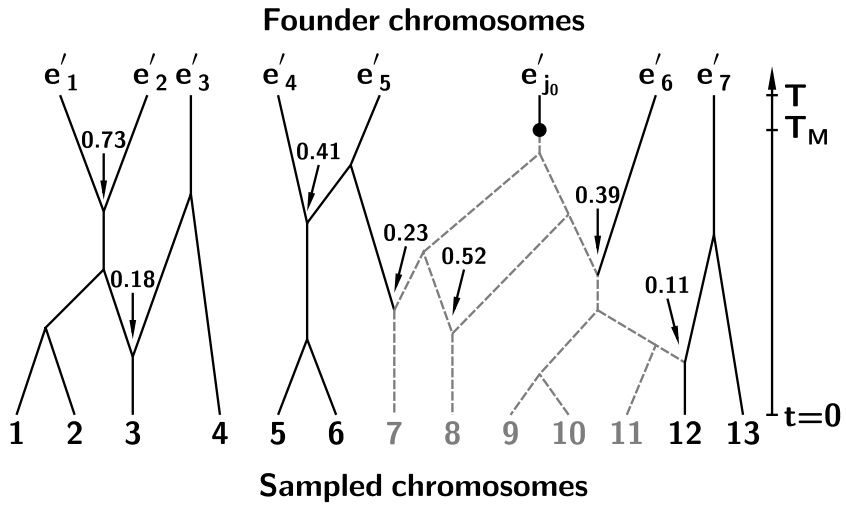

Figure 1: Ancestral recombination graph for $n = 13$ chromosomes. Lineages carrying the mutation, positioned at $\tau = 0.36$, are displayed with dashed gray lines. $M_s = \{7, \ldots, 11\}$ is the sampled mutated chromosomes, whereas the other 8 sampled chromosomes do not carry the mutation. The time-scale on the right measures time backwards from today's sample until the founder generation.

To include ascertainment in the analysis the population is split into two subpopulations M and U, which are the union of mutated and unmutated chromosomes respectively, *at all time points* $0 \leq t \leq T$. Write $e \overset{t}{\in} M$ to indicate that the chromosome of lineage e at time t is mutated. Notice that all chromosomes of e belong to the same subpopulation, unless e contains the mutated chromosome at time $T_M = G_M/G_M = 1$. In that case we may have $e \overset{t}{\in} M$ for $t \leq T_M$ and $e \overset{t}{\in} U$ for $t > T_M$.

The dynamics of the ARG under H_1^x is described by a Wright-Fisher model in reversed (and rescaled) time, with varying population size. That is, $N(\cdot)$, $N_M(\cdot)$ and $N_U(\cdot)$ are considered fixed and non-stochastic. If the unmutated population is large at all time points, two unmutated edges at time t coalesce at rate

$$\mu_U(t) = G_M/N_U(t), \quad 0 \leq t \leq T,$$

accounting for that time is speeded up by a factor G_M. Two mutated edges

coalesce at rate

$$\mu_M(t) = \begin{cases} G_M \log(1/(1-1/N_M(t))), & 0 \le t < T_M, \\ \infty, & t = T_M. \end{cases}$$

Since $N_M(t) = 0$ for $t > T_M$, $\mu_M(t)$ is then undefined. We put $\mu_M(T_M) = \infty$ to describe formally that all remaining mutated edges have to coalesce at time T_M. Notice that $\mu_M(t) \approx G_M/N_M(t)$ when $N_M(t)$ is large. However, we cannot use this approximation for t close to T_M, since then $N_M(t)$ is small. Instead we utilize that the probability of no coalescence per generation is $(1 - 1/N_M(t))$, take the logarithm and change sign to get a rate, and finally speed up the rate by a factor G_M.

To describe the evolvement of the ARG in time, we let $n(t)$ be the number of lineages (edges) at time t, so that $n = n(0)$ is the number of sampled chromosomes. Each coalescence/recombination event decreases/increases $n(t)$ by one. Let $n_M(t)$ and $n_U(t) = n(t) - n_M(t)$ be the number of mutated and unmutated lineages at time t, with $n_M = n_M(0)$ and $n_U = n_U(0)$.

Assume that recombinations occur with probability r per chromosome and generation. Since the chromosome region of interest is typically very small, so is r. For this reason, we use the rescaled recombination rate

$$\rho = G_M r$$

rather than r. Given that a recombination occurs, it has density π on $[0, 1]$, where $\pi \equiv 1$ is uniform on $[0, 1]$ if genetic distance along the chromosome is used. However, $\pi(\cdot)$ often varies when physical distance is used, to reflect varying recombination rate or recombination hot-spots.

In our definition of an ascertainment corrected ARG, we formalize the constraints that subpopulations M and U put on segregation by subdividing jumps of the ARG at time t into four categories:

1. Coalescence between two mutated edges at rate $\binom{n_M(t)}{2}\mu_M(t)$. Given such an event, choose uniformly a pair $e_1, e_2 \overset{t}{\in} M$ to coalesce among all $\binom{n_M(t)}{2}$ possible. For the coalesced edge e we put $e \overset{t}{\in} M$ so that $n_M(t+) = n_M(t-) - 1$ and $n_U(t+) = n_U(t-)$.

2. Coalescence between two unmutated edges at rate $\binom{n_U(t)}{2}\mu_U(t)$. Given such an event, choose uniformly a pair $e_1, e_2 \overset{t}{\in} U$ among all $\binom{n_U(t)}{2}$ possible. For the coalesced edge e we put $e \overset{t}{\in} U$ so that $n_U(t+) = n_U(t-) - 1$ and $n_M(t+) = n_M(t-)$.

3. A mutated chromosome recombines at rate $n_M(t)\rho$. Given such an event, choose uniformly $e \overset{t}{\in} M$ among all $n_M(t)$ mutated edges and recombination point $x \sim \pi$. Let e_1 and e_2 be the two parental lines of e which have passed on material $[0, x)$ and $[x, 1]$ respectively to e. Then

$$P(e_1 \overset{t}{\in} M | e \overset{t}{\in} M) = \begin{cases} p(t), & x < \tau, \\ 1, & x \geq \tau. \end{cases}$$

and

$$P(e_2 \overset{t}{\in} M | e \overset{t}{\in} M) = \begin{cases} 1, & x < \tau, \\ p(t), & x \geq \tau, \end{cases}$$

where $p(t) = N_M(t)/N(t)$ is the proportion of mutated chromosomes in the population at time t. Hence $n_M(t+) = n_M(t-) + 1$, $n_U(t+) = n_U(t-)$ with probability $p(t)$ and $n_M(t+) = n_M(t-)$, $n_U(t+) = n_U(t-)+1$ with probability $1 - p(t)$.

4. An unmutated chromosome recombines at rate $n_U(t)\rho$. Given such an event, choose uniformly $e \overset{t}{\in} U$ among all $n_U(t)$ unmutated edges and recombination point $x \sim \pi$. Let e_1 and e_2 be the two parental lines of e which have passed on material $[0, x)$ and $[x, 1]$ respectively to e. Then

$$P(e_1 \overset{t}{\in} M | e \overset{t}{\in} U) = \begin{cases} p(t), & x < \tau, \\ 0, & x \geq \tau. \end{cases}$$

and

$$P(e_2 \overset{t}{\in} M | e \overset{t}{\in} U) = \begin{cases} 0, & x < \tau, \\ p(t), & x \geq \tau. \end{cases}$$

Hence $n_M(t+) = n_M(t-) + 1$, $n_U(t+) = n_U(t-)$ with probability $p(t)$ and $n_M(t+) = n_M(t-)$, $n_U(t+) = n_U(t-)+1$ with probability $1-p(t)$.

The initial state of $A(0)$ of the ARG is determined by the set of n_M *m*utated *s*ampled chromosomes, which we denote by

$$M_s = \{i;\, 1 \leq i \leq n \text{ and } i \in M\}.$$

The further evolvement of $\mathcal{A}$ under H_1^x up to time T is fully determined by the four kinds of transition events described above. At time $t = T = G/G_M$, let $e'_1, \ldots, e'_{n'}$ denote the edges of $\mathcal{A}$, numbered from left to right. We think of these edges as n' founder chromosomes whose genetic material is segregated/gene dropped down to the present generation, and possibly mutated. Let $h'_j = (h'_{j1}, \ldots, h'_{jK})$ denote the haplotype of the j^{th} founder chromosome and $\boldsymbol{h}' = (h'_1, \ldots, h'_{n'})$ the collection of all founder haplotypes.

The distribution of $\mathcal{A}$ under H_0 is equivalent to putting $N_M(\cdot) \equiv 0$ above, i.e. $M_s = \emptyset$, $p(t) \equiv 0$ and $n_M(t) \equiv 0$.

Mutations that are selectively neutral and have no influence on the phenotype can be positioned along the edges of the ARG as follows: Assume mutation probability u_k per generation and chromosome at locus x_k and put $\theta_k = G_M u_k$. Then mutations are positioned along the edges of $\mathcal{T}(x_1), \ldots, \mathcal{T}(x_K)$ according to K independent Poisson processes with intensities $\theta_1, \ldots, \theta_K$. In addition, the state space of alleles and transition probabilities between alleles must be defined at each marker locus.

4 Retrospective Likelihood

In order to give an expression for the retrospective likelihood (1), we think of $\boldsymbol{h}$ obtained by segregating haplotypes $\boldsymbol{h}'$ of the founder generation at all marker loci according to $\mathcal{A}$, superimposed by neutral mutations. This we write as

$$L(x) = \sum_{\boldsymbol{h},\boldsymbol{h}',\mathcal{A},M_s} P(\boldsymbol{g}|\boldsymbol{h}) P(\boldsymbol{h}|\mathcal{A},\boldsymbol{h}') P(\boldsymbol{h}'|n') P_x(\mathcal{A}|M_s) P(M_s|\boldsymbol{Y}). \qquad (5)$$

The term $P_x(\mathcal{A}|M_s)$ is the probability of the ARG with initial condition M_s and

$$P(\boldsymbol{g}|\boldsymbol{h}) = \prod_{v=1}^{m} P(g_v|h_{2v-1}, h_{2v}) = \prod_{v=1}^{m} 1_{\{g_v \sim (h_{2v-1}, h_{2v})\}},$$

where $g_v \sim (h_{2v-1}, h_{2v})$ means that the genotypes of v at all K loci are consistent with the corresponding haplotype vectors.

The term $P(M_s|\boldsymbol{Y})$ in (5) only depends on genetic model parameters of the disease. Define penetrance parameters

$$\psi_{vj} = P(Y_v | v \text{ has } j \text{ disease alleles } B),$$

$v = 1, \ldots, m$, $j = 0, 1, 2$, and let p_j be the probability that a randomly sampled genotype at time 0 has j B-alleles. Conditional on $\boldsymbol{Y}$, the mutation status indicators $(2v - 1 \in M_s, 2v \in M_s)_{v=1}^m$ are independent pairs of binary random variables, so that

$$P(M_s|\boldsymbol{Y}) = \prod_{v=1}^{m} P(M_s \cap \{2v-1, 2v\} | Y_v) \qquad (6)$$

with

$$\begin{array}{rcl} P(2v-1 \in M_s, 2v \in M_s|Y_v) & = & \psi_{v2}p_2/S_v \\ P(2v-1 \notin M_s, 2v \in M_s|Y_v) & = & \psi_{v1}p_1/(2S_v) \\ P(2v-1 \in M_s, 2v \notin M_s|Y_v) & = & \psi_{v1}p_1/(2S_v) \\ P(2v-1 \notin M_s, 2v \notin M_s|Y_v) & = & \psi_{v0}p_0/S_v \end{array} \tag{7}$$

and $S_v = \psi_{v0}p_0+\psi_{v1}p_1+\psi_{v2}p_2$. In general we have the constraint $p_2+0.5p_1 = p$, where $p = p(0) = N_M/N$ is the disease allele frequency. Under Hardy-Weinberg equilibrium $p_0 = (1-p)^2$, $p_1 = 2p(1-p)$ and $p_2 = p^2$.
Let f denote frequencies of founder haplotypes defined over marker loci $x_1, \ldots, x_K$. Assuming founder haplotypes are independent we get

$$P(\boldsymbol{h}'|n') = \prod_{j=1}^{n'} f(h'_j). \tag{8}$$

Let $A(T)$ be the state of the ARG at time T. It carries information about which founders that have segregated chromosome segments down to the sample. The information that $A(T)$ carries about segregation at marker loci $x_1, \ldots, x_K$ can be represented as a decomposition of $\Omega = \{1, \ldots, n\} \times \{1, \ldots, K\}$ into n' disjoint sets $D_1, \ldots, D_{n'}$, where $(i,k) \in D_j$ iff e'_j is ancestral to i at locus x_k. Figure 2 displays the decomposition $\{D_j\}_{j=1}^{n'}$ corresponding to the ARG of Figure 1. Identical by descent (IBD) relative to the founder population is defined so that all alleles h_{ik}, $(i,k) \in D_j$ are IBD. Further, let E_j be the projection of D_j onto $\{1, \ldots, K\}$, i.e. the set of k such that e'_j is ancestral to at least one chromosome i at locus x_k. Write $\boldsymbol{h}_C = \{h_{ik}, (i,k) \in C\}$ for any subset $C \subset \Omega$. Then

$$P(\boldsymbol{h}|\mathcal{A}, \boldsymbol{h}') = \prod_{j=1}^{n'} \prod_{k \in E_j} P(\boldsymbol{h}_{D_{jk}}|\mathcal{T}(x_k), h'_{jk}), \tag{9}$$

where $D_{jk} = D_j \cap (\{1, \ldots, n\} \times \{k\})$ is the k^{th} column of D_j. Notice that $\mathcal{T}(x_k)$ consists of a number of disjoint subtrees. Each such subtree has a root at e'_j for some j such that $k \in E_j$ and leaves at all i such that $(i,k) \in D_{jk}$. The term $P(\boldsymbol{h}_{D_{jk}}|\mathcal{T}(x_k), h'_{jk})$ in (9) depends on neutral mutations at locus x_k along the subtree of $\mathcal{T}(x_k)$ with root e'_j.

5 Model Simplifications

The retrospective likelihood (5) is general, but involves summation over all ARGs consistent with data. Due to the large space of possible ARGs the computation is very computer intensive, and has to be carried out by Monte

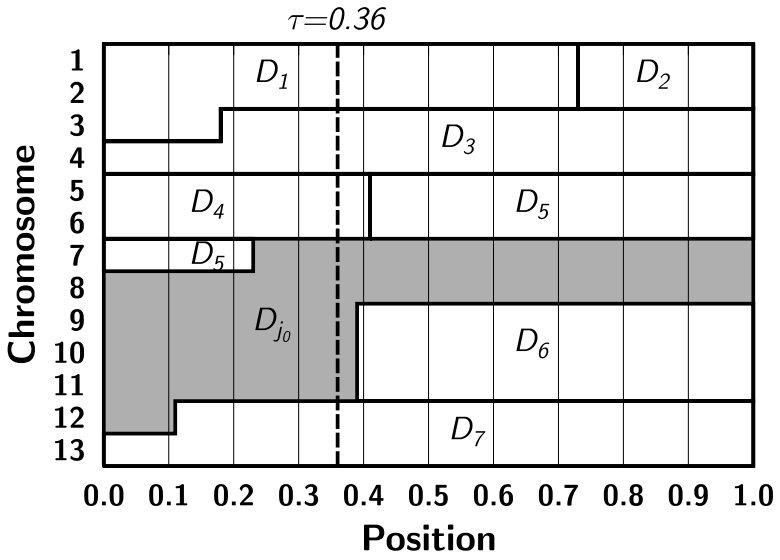

Figure 2: IBD regions for the ARG of Figure 1. The mutated region D_{j_0} is displayed in gray. The location of the disease mutation $\tau = 0.36$, as well as 11 marker positions, $x_1 = 0,\ x_1 = 0.1, \ldots,\ x_{11} = 1$ are displayed with vertical lines.

Carlo. Still, a feasible algorithm would require development of efficient sampling algorithms of $\mathcal{A}$ given $\boldsymbol{g}$ and $\boldsymbol{Y}$, which avoids genealogies that are not consistent with data. As an alternative to Monte Carlo, we will in this section instead consider approximations and model simplifications that make the exact computation feasible. We assume:

i. The mutated population is a small fraction of the total population at all time points, i.e. $\max_{0 \le t \le T_M} p(t) \ll 1$. In particular this implies a rare disease allele $p \ll 1$. As a consequence, new mutated edges are never created in the ARG when going backwards in time and $n_M(\cdot)$ is non-increasing.

ii. No coalescence of unmutated chromosomes, since G_M is too small in comparison to the size of the unmutated population. The formal criterion is $\int_0^T \mu_U(t)dt \ll 1$.

iii. Star topology of the subtree of $\mathcal{T}(\tau)$ with n_M mutated edges. This corresponds to $\mu_M(t) \ll 1$ for $t \in [0, T_M - \varepsilon]$ and some small $\varepsilon > 0$. Hence all n_M mutated lines at locus τ coalesce simultaneously at time T_M.

iv. Founder population at time of disease mutation, i.e. $T = T_M$. This is in order to simplify analysis and implies that one of the n' founder haplotypes is the mutated disease chromosome, so that $n_M(T) = 1$ and $n_U(T) = n' - 1$.

v. There is linkage equilibrium (LE) in the founder population, i.e. $f(h'_j) = \prod_{k=1}^{K} f_k(h'_{jk})$, $j = 1, \ldots, n'$, with f_k denoting founder allele frequency at locus x_k.

vi. All markers are bi-allelic SNPs with mutations occurring along all edges of $\mathcal{T}(x_k)$ according to a Markov process in continuous time with intensity matrix

$$\begin{pmatrix} -\theta_k & \theta_k \\ \theta_k & -\theta_k \end{pmatrix}.$$

Conditions i-iv are further discussed in the appendix for the exponential growth model (4). Figure 3 shows an ARG with 8 chromosomes, satisfying the approximations. The subpopulations corresponding to a disease locus at $\tau = 0.36$ are displayed.
Since no lineages coalesce before time $T = T_M$, the decomposition $\{D_j\}_{j=1}^{n'}$ of Ω into disjoint IBD regions carries all necessary information of $\mathcal{A}$ for computing $L(x)$, whenever i–vi hold. Thus

$$P(\boldsymbol{h}|\mathcal{A}, \boldsymbol{h}') = P(\boldsymbol{h}|D_1, \ldots, D_{n'}, \boldsymbol{h}') = P(\boldsymbol{h}_{D_{j0}}|h'_{j0}) \prod_{j \neq j0} P(\boldsymbol{h}_{D_j}|h'_j) \qquad (10)$$

Figure 4 shows $\{D_j\}_{j=1}^{n'}$ for the ARG of Figure 3, where the disease mutation is located at $\tau = 0.36$. There is one IBD region D_{j_0} of mutated chromosomes and the remaining $n' - 1$ regions D_j contain unmutated chromosomes. Since there are no coalescence events of the ARG when $t \in [0, T_M]$ along lineages starting at $i \notin M_s$, none of the regions D_j, $j \neq j_0$ extends over more than one single row $(i, \cdot) = \{i\} \times \{1, \ldots, K\}$. However, i may have several ancestors in the founder generation due to recombination events, so that each $(i, \cdot)$ may contain several D_j.

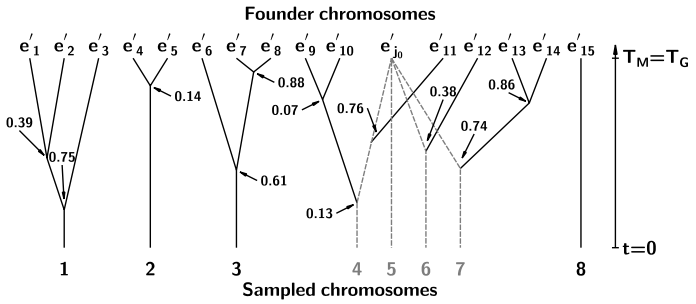

Figure 3: Ancestral recombination graph for $n = 8$ chromosomes from $n' = 16$ founders. Mutated lineages are displayed with dashed lines. $M_s = \{4, \ldots, 7\}$, whereas the other 4 sampled chromosomes are unmutated. The time-scale on the right measures time backwards from today's sample until the founder generation.

In the sequel, we simplify notation by omitting index j_0, writing

$$\begin{aligned} D &= D_{j_0}, \\ h' &= h'_{j_0} = (h'_{j_0 1}, \ldots, h'_{j_0 K}) \end{aligned}$$

Using (10) and summing over $\boldsymbol{h}'_{(-j_0)} = \{h'_j;\, j \neq j_0\}$, $\{D_j;\, j \neq j_0\}$ and M_s, the likelihood (5) becomes

$$L(x) = \sum_{\boldsymbol{h}, h', D} P(\boldsymbol{g}|\boldsymbol{h}) P(\boldsymbol{h}|D, h') f(h') P_x(D|\boldsymbol{Y}). \tag{11}$$

The region D of mutated chromosomes can be defined as follows: Assume H_1^x, with $x_{k_0} \leq x < x_{k_0+1}$ for some $k_0 = 0, 1, 2, \ldots, K$. (We put $x_0 = -\infty$ and $x_{K+1} = \infty$ to make k_0 well defined even when $x < x_1$ or $x \geq x_K$.) For each $i \in M_s$, consider the i-lineage of $\mathcal{T}(x)$ from $t = 0$ back to $t = T_M$. It is a union of several edges along the ARG. Each junction between two such edges corresponds to a recombination event. Let X_i^- and X_i^+ be the recombination points to the left and right of x that are closest to x. (If there are no recombination points to the left of x we put $X_i^- = -\infty$ and similarly $X_i^+ = \infty$ if there are no recombination points to the right of x.) Then

$$\begin{aligned} D = \{&(i,k) \in \Omega,\ i \in M_s \text{ and either } x_k \geq X_i^- \text{ for } k \leq k_0 \\ &\text{or } x_k < X_i^+ \text{ for } k \geq k_0 + 1\}. \end{aligned} \tag{12}$$

Thus $P_x(D|\boldsymbol{Y}) = \sum_{M_s} P_x(D|M_s) P(M_s|\boldsymbol{Y})$, the conditional probability of the mutated IBD region D given phenotypes, involves the genetic model

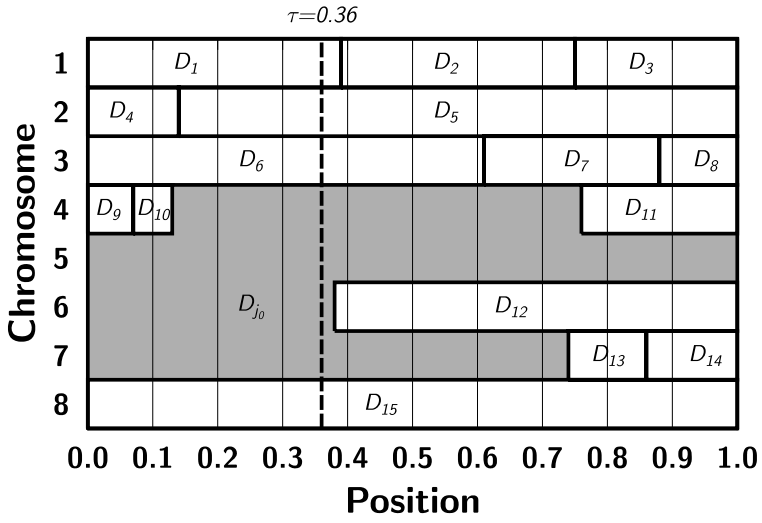

Figure 4: IBD regions for the ARG of Figure 3, i.e. with model approximations. The mutated region D_{j_0} is displayed in gray. Note that no other IBD regions extend over more that one line. The location of the disease mutation $\tau = 0.36$, as well as 11 marker positions, $x_1 = 0,\ x_1 = 0.1, \ldots,\ x_{11} = 1$ are displayed with vertical lines.

through $P(M_s|\boldsymbol{Y})$ in (6), and the disease locus parameter x and the recombination parameters ρ and $\pi(\cdot)$ through $P_x(D|M_s)$ in (12) and (A.2). The term $f(h')$ depends on founder haplotype frequencies and $P(\boldsymbol{h}|D, h')$ involves both founder haplotype frequencies and mutation rate parameters.

Conditions v-vi imply

$$P(\boldsymbol{h}|D, h') = \prod_{(i,k)\in D} P(h_{ik}|h'_{j_0k}) \cdot \prod_{(i,k)\in\Omega\setminus D} \tilde{f}_k(h_{ik}) \tag{13}$$

where

$$P(h_{ik}|h'_{j_0k}) = (1-q_k)^{\{h_{ik}=h'_{j_0k}\}} q_k^{\{h_{ik}\neq h'_{j_0k}\}}, \tag{14}$$

$q_k = (1-\exp(-2\theta_k))/2 \approx \theta_k$ is the probability that a founder allele at time $t = 1$ mutates an odd number of times down to $t = 0$ and

$$\tilde{f}_k(a) = (1-q_k)f_k(a) + q_k f_k(1-a), \quad a = 0, 1, \tag{15}$$

is the allele frequency of a at locus x_k in today's generation. If θ_k is small, $\tilde{f}_k(a) \approx f_k(a)$.

6 LOD Score Computation and Approximation

Based on (11)-(15), we will compute the likelihood ratio $\mathrm{LR}(x)$ and the associated LOD score (2). The first step is to calculate the likelihood (11) when $x = \infty$. Since $P_\infty(D = \emptyset|\boldsymbol{Y}) = 1$, we get

$$\begin{array}{rcl} L(\infty) & = & \sum_{\boldsymbol{h}} P(\boldsymbol{g}|\boldsymbol{h}) \prod_{(i,k)\in\Omega} \tilde{f}_k(h_{ik}) \\ & = & n_{\boldsymbol{g}} \cdot \prod_{(i,k)\in\Omega} \tilde{f}_k(h_{ik}), \end{array} \tag{16}$$

where $n_{\boldsymbol{g}} = |\{\boldsymbol{h};\, \boldsymbol{h} \sim \boldsymbol{g}\}|$ is the number of haplotype configurations consistent with genotype data, i.e. the number of $\boldsymbol{h}$ such that $P(\boldsymbol{g}|\boldsymbol{h}) = 1$. In the last step of (16) we used that $\prod_{(i,k)\in\Omega} \tilde{f}_k(h_{ik})$ is the same for any $\boldsymbol{h} \sim \boldsymbol{g}$.

Taking the ratio of (11) and (16) we get

$$\mathrm{LR}(x) = \sum_{h',D} \mathrm{LR}(h', D) f(h') P_x(D|\boldsymbol{Y}), \tag{17}$$

where $\mathrm{LR}(h', D) = P(\boldsymbol{g}|h', D) / \left(n_{\boldsymbol{g}} \prod_{(i,k)\in\Omega} \tilde{f}_k(h_{ik})\right)$ is the likelihood ratio obtained when conditioning on missing data (h', D). It turns out that a very explicit expression for $\mathrm{LR}(h', D)$ can be obtained. To this end, introduce $H \subset \Omega$ as the the set of heterozygous sites (i, k), i.e.

$$H = \cup_{k=1}^{K} \cup_{v=1}^{m} H_{vk},$$

where H_{vk} is empty if g_{vk} is homozygous ($h_{2v-1,k} = h_{2v,k}$) and $H_{vk} = \{(2v-1, k), (2v, k)\}$ if g_{vk} is heterozygous ($h_{2v-1,k} \neq h_{2v,k}$). The expression for $\mathrm{LR}(h', D)$ is obtained by taking the ratio of the right-hand sides of (13) and (16) and summing over all $\boldsymbol{h} \sim \boldsymbol{g}$. This sum involves switching alleles independently of all heterozygous genotypes, i.e. switching alleles within each nonempty H_{vk}, yielding $n_{\boldsymbol{g}} = 2^{|H|/2}$ terms. The contribution to $\mathrm{LR}(h', D)$ is independent for all genotypes g_{vk} and depends on zygosity of g_{vk} as well as how $\{(2v-1, k), (2v, k)\}$ intersects with D. Let H_k consist of those heterozygous sites that belong to the k^{th} column of D but the homologous site (i.e., the member of the same H_{vk}) does not. Then the k^{th} column of D has $n_{kH} + n_{k0} + n_{k1}$ elements, where $n_{kH} = |H_k|$, $n_{k0} = |\{i;\, (i, k) \in D \setminus H_k, h_{ik} = 0\}|$ and $n_{k1} = |\{i;\, (i, k) \in D \setminus H_k, h_{ik} = 1\}|$ and

$$\begin{array}{rcl} \mathrm{LR}(h', D) & = & \prod_{k=1}^{K} \Big(\left(P(0|h'_{j_0k})/\tilde{f}_k(0)\right)^{n_{k0}} \left(P(1|h'_{j_0k})/\tilde{f}_k(1)\right)^{n_{k1}} \\ & & \cdot \left(0.5P(0|h'_{j_0k})/\tilde{f}_k(0) + 0.5P(1|h'_{j_0k})/\tilde{f}_k(1)\right)^{n_{kH}} \Big). \end{array} \tag{18}$$

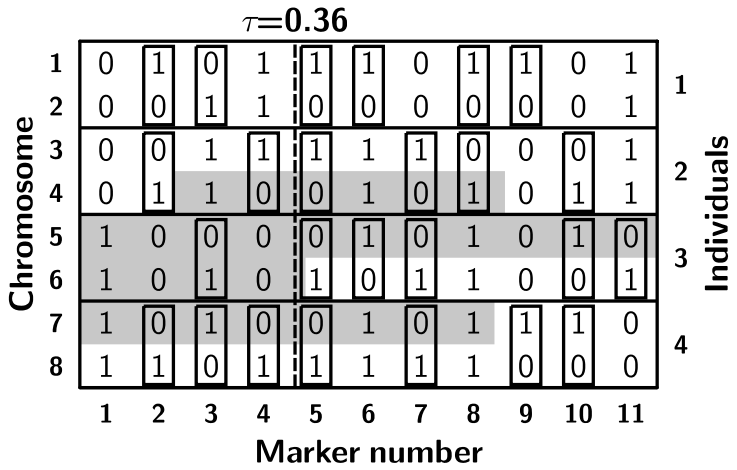

Figure 5: An example of marker haplotypes for the sample in Figure 3 and 4, to illustrate H, the set of heterozygous sites, and H_k, the intersection between $D = D_{j_0}$ and column k in H. The mutated region D is displayed in gray (note that a mutation seems to have taken place at marker 3 in chromosome 5). The set of heterozygous sites, H, is the union of the marked boxes. For each marker $1 \leq k \leq K$, the set H_k consists of those chromosomes at heterozygous sites that belong to the k^{th} column of D, but where the homologous site does not. Thus in this example $H_1 = \emptyset$, $H_2 = \{7\}$, $H_3 = \{7\}$, $H_4 = \{4, 7\}$, $H_5 = \{4, 5, 7\}$, $H_6 = \{5\}$, $H_7 = \{4, 5, 7\}$, $H_8 = \{4\}$, $H_9 = \emptyset$, $H_{10} = \{5\}$, $H_{11} = \{5\}$. Further it follows that $n_{3H} = 1$, $n_{30} = 1$ and $n_{31} = 2$, etc.

which are computer intensive, but much less so than (17).

6.1 Conditioning on Founder Haplotypes

In this approach we sum out D in (17) and write

$$\mathrm{LR}(x) = \sum_{h'} \mathrm{LR}(x; h') f(h'), \tag{19}$$

where $\mathrm{LR}(x; h') = P_x(\boldsymbol{g}|h', \boldsymbol{Y}) / \left(n_{\boldsymbol{g}} \prod_{(i,k)\in\Omega} \tilde{f}_k(h_{ik}) \right)$ is the likelihood ratio when conditioning on missing data h'. Let $R_v = D \cap (\{2v-1, 2v\} \times \{1, \ldots, K\})$ denote the set of mutated sites (i,k) for Individual v. Then (17)-(18) imply

$$\mathrm{LR}(x; h') = \prod_{v=1}^{m} \sum_{R_v} \mathrm{LR}(h', R_v) P_x(R_v|Y_v) \tag{20}$$

where $\mathrm{LR}(h'; R_v)$ is the likelihood ratio obtained when conditioning on hidden data (h', R_v), i.e. replacing D by R_v in (18). The crucial point is that conditionally on h' and $\boldsymbol{Y}$, the rows of $\boldsymbol{g}$ are independent and hence LR can be written as a product of m terms. It is shown in the appendix that each term of the outer product can be calculated with $O(K)$ operations, using a recursive Hidden Markov Model (HMM) algorithm. Hence the total complexity is $O(mK2^K)$ for evaluating $\mathrm{LR}(x)$. This is a marked improvement over direct summation over h' and D, but still un-feasible for large K. For large K, we may use a sliding window of $l < K$ marker loci. The window width l is chosen to make the computational complexity $O(ml2^l K)$ feasible. To obtain p-values for the test, a permutation algorithm was proposed in (3). In general permutation tests are very computationally intensive, which constrict their practical applicability for tests that are already computationally demanding, such as ours. In the general setting, the test quantity must be calculated for each of the Q random permutations, which would give computational complexity $O(mK2^K Q)$. Since Q must be large, typically tens or hundreds of thousands, this is not feasible. However, in the case of binary phenotypes, we propose a procedure for the permutation testing which reduce the computational demand. The algorithm exploits that $\sum_{R_v} \mathrm{LR}(h', R_v) P_x(R_v|Y_{\gamma(v)})$ is the same for all permutations where $Y_{\gamma(v)} = 1$, and similarly for all permutations where $Y_{\gamma(v)} = 0$. Thus, for each individual v the HMM must only be calculated twice, to obtain $\sum_{R_v} \mathrm{LR}(h', R_v) P_x(R_v|Y_v = 1)$ and $\sum_{R_v} \mathrm{LR}(h', R_v) P_x(R_v|Y_v = 0)$ respectively.
To obtain p-values the summation over h' and multiplication over v in (19) and (20) must be carried out for each of the Q permutations. The total

complexity is thus $O(m2^K(2K+Q))$. Since typically $Q >> K$ the total complexity including permutation testing is $O(m2^K Q)$.

To estimate Z_{max}, LOD score is calculated at several positions $\tilde{x}_i$, $i = 1, \ldots, N_x$ within the interval [0, 1], and $Z_{max} = \max_{i=1,\ldots,N_x} Z(\tilde{x}_i)$. As LOD score is calculated separately at each position, the total complexity is $O(m2^K Q N_x)$.

6.2 Conditioning on IBD Regions

In this approach we sum out h' in (17) and write

$$\mathrm{LR}(x) = \sum_D \mathrm{LR}(D) P_x(D|\boldsymbol{Y}), \tag{21}$$

where $\mathrm{LR}(D) = P(\boldsymbol{g}|D)/\left(n_{\boldsymbol{g}} \prod_{(i,k)\in\Omega} \tilde{f}_k(h_{ik})\right)$ is the likelihood ratio obtained when conditioning on missing data D. This yields

$$\begin{array}{rcl} \mathrm{LR}(D) & = & \prod_{k=1}^K \Big(a_k^{n_{kH}}((1-q_k)/\tilde{f}_k(0))^{n_{k0}}(q_k/\tilde{f}_k(1))^{n_{k1}} f_k(0) \\ & & \quad + b_k^{n_{kH}}(q_k/\tilde{f}_k(0))^{n_{k0}}((1-q_k)/\tilde{f}_k(1))^{n_{k1}} f_k(1)\Big), \end{array} \tag{22}$$

where $a_k = 0.5(1-q_k)/\tilde{f}_k(0) + 0.5q_k/\tilde{f}_k(1)$ and similarly $b_k = 0.5q_k/\tilde{f}_k(0) + 0.5(1-q_k)/\tilde{f}_k(1)$.

In the appendix, we describe a HMM algorithm for evaluating (21) with complexity $O(K2^{2m})$. This is un-feasible for all but very small m, so we propose using a pseudo likelihood

$$\mathrm{PL}(x) = \prod_{V\in\mathcal{V}} L(x;V), \tag{23}$$

where $L(x;V)$ is the retrospective likelihood using only individuals from $V \subset \{1,\ldots,m\}$ and $\mathcal{V}$ a given collection of subsets V. The pseudo likelihood ratio and pseudo LOD score obtained from (23) are

$$\mathrm{PLR}(x) = \prod_{V\in\mathcal{V}} \mathrm{LR}(x;V)$$

and

$$\begin{array}{rcl} \mathrm{PZ}(x) & = & \frac{1}{|\mathcal{V}|}\log_{10}(\mathrm{PLR}(x)) \\ & = & \frac{1}{|\mathcal{V}|}\sum_{V\in\mathcal{V}} Z(x;V). \end{array}$$

For instance, if $\mathcal{V}$ contains all subsets of $\{1,\ldots,m\}$ of size m_0, we get computational complexity $O(Km^{m_0}2^{2m_0})$, which, for values of m of practical interest, is feasible for m_0 at most 2.

When a permutation test is used to calculate the p-values the procedure would in general require $O(Km^{m_0}2^{2m_0}Q)$ operations for the PLOD score. For $m_0 = 2$ this is $O(Km^2 2^4 Q) = O(Km^2 Q)$. However, for binary phenotypes an effective algorithm can be developed, just as for the LOD score. The basis of this algorithm is that $\mathrm{LR}(x;V) = \sum_{D_V} \mathrm{LR}(D_V) P_x(D_V|\boldsymbol{Y}_V)$ is constant for all permutations with the same $\boldsymbol{Y}_V$. Here D_V is notation for which markers that are inherited IBD (from the mutated founder) for the individuals *within* subset V. As $\mathcal{V}$ consists of subsets of size 2, $\boldsymbol{Y}_V$ can take only four possible values, $\boldsymbol{Y}_V = (0\ 0)$, $\boldsymbol{Y}_V = (0\ 1)$, $\boldsymbol{Y}_V = (1\ 0)$ or $\boldsymbol{Y}_V = (1\ 1)$. Thus $\mathrm{LR}(x;V)$ must be calculated for each of these four cases, and for each permutation only the multiplication over all subsets in $\mathcal{V}$ remains. The complexity is thus $O(m^2(4K2^4 + Q))$. Since typically $Q \gg 2^6 K$ the complexity becomes $O(m^2 Q)$. To estimate $\mathrm{PZ}_{\max}$, PLOD score is calculated at several positions $\tilde{x}_i,\ i = 1, \ldots, N_x$ within the interval $[0,\ 1]$, and $\mathrm{PZ}_{\max} = \max_{i=1,\ldots,N_x} \mathrm{PZ}(\tilde{x}_i)$. As the PLOD score is calculated separately at each position, the total complexity is $O(m^2 Q N_x)$.

6.3 Software

The algorithms for simulation and calculation of the LOD and PLOD scores have been coded in Matlab. The algorithms, with inbuilt documentation, are available after request from the authors.

7 Simulation study

To evaluate the performance of the proposed LOD and PLOD scores, a small simulation study is presented.

As previously pointed out, the retrospective ARG is a powerful tool for simulation of case-control samples. For a prescribed number of cases and controls, the mutational status for each of a person's two alleles at the disease locus is simulated conditional on the person's disease status, according to (6). By simulation of the four different events 1–4 on Page 8, superimposed by neutral mutations, the marker alleles are obtained. The following simulations are obtained under the simplifications (i)–(vi) in Section 5.

As an example of a commonly used genetic model, we account for simulations with multiplicative penetrance and binary phenotype. With genotype relative risk ratio λ, we have $\psi_1/\psi_0 = \psi_2/\psi_1 = \lambda$, where ψ_j is the probability that an individual with j disease alleles becomes affected. (Then $\psi_{vj} = \psi_j$ for all cases ($Y_v = 1$) and $\psi_{vj} = 1 - \psi_j$ for the controls ($Y_v = 0$).) Further HW equilibrium was assumed, i.e. genotype frequencies $p_0 = (1-p)^2$,

$p_1 = 2p(1-p)$ and $p_2 = p^2$. The markers were equispaced in the interval $[0,1]$ (with $x_1 = 0, \ldots, x_K = 1$), with minor marker allele frequency $f_k = 0.5$ at all markers $k = 1, \ldots, K$. From the founder generation until today, the marker mutation rate was $q_k = q = 0.001$ and recombination rate in the interval was $\rho = 1.5$. In all simulations the disease locus was positioned at $\tau = 0.36$, which was not a marker position. All parameter values can be chosen arbitrarily, although their values affect the power to detect association. Considering mutations that arose typically some hundred generations ago, the mutation rate 0.001 per marker is unrealistically high for SNPs, but still does not undermine the performance of our LOD score. On the other hand, the marker allele frequencies $f_k = 0.5$ are unrealistical to our favour. The accompanying decrease in sample size, that is made possible, is welcome for the computer demanding studies of power that we present here. However it does not change the fundamental behaviour of the LOD score, compared to arbitrary marker allele frequencies. We further test our algorithms for parameter values that do not fulfil all conditions of the approximation. In particular, the disease allele frequency is too high in the first simulation, and in that way more similar to what is assumed in real studies. (To pick up associations for diseases with weak penetrance would need unrealistically large samples if the disease allele frequency was very low.)
Each simulated data set was analyzed with the LOD score (19) and/or PLOD score (23), the latter with subsets of $m_0 = 2$ individuals. (For some data sets either of the methods was un-feasible due to the computational demand). To evaluate the performance, the p-value of the test statistic Z_{max} was found by permutation testing (3).

7.1 LOD score and PLOD score

In three independent samples of 200 cases and 200 controls at $K = 5$ markers, the disease allele frequency was $p = 0.2$, relative risk ratio $\lambda = 3$, and prevalence 0.001. Figure 6 displays the LOD and PLOD scores of the three data sets, each calculated at $N_x = 20$ equidistant locations interior of $[0,1]$. To be able to detect associations with a p-value as small as 10^{-4}, 100000 permutations were performed.
For the LOD score, the association is very clearly picked up by Z_{max}, and further it is clear that the largest $Z(x)$ is found close to the true maximum $\tau = 0.36$. Although the shape of the PLOD score curve is similar to that of LOD score, with its maximum close to $\tau = 0.36$, the p-values are considerably lower for PLOD. The p-value calculations further show that permutation testing is necessary, since the asymptotic χ^2-approximation (in which case LOD=3 corresponds to p-value 0.0002, which is commonly used to establish

linkage) is not valid neither for LOD nor PLOD .

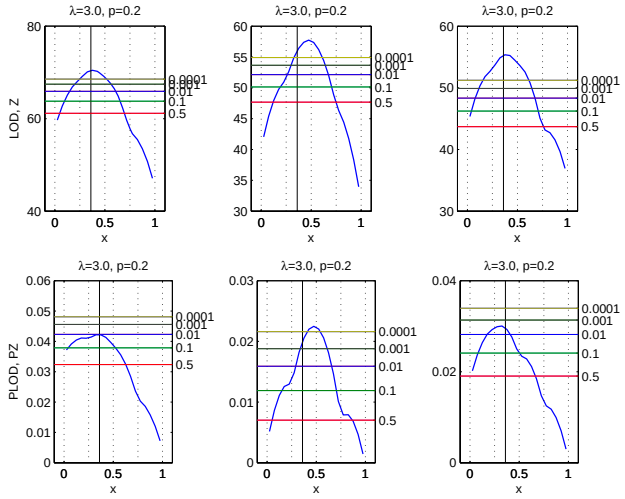

Figure 6: LOD and PLOD score calculated for three simulated data sets, at $N_x = 20$ positions along $[0, 1]$. The genetic model is multiplicative penetrance with relative risk 3, disease allele frequency 0.2, marker allele frequency $f_k(1) = 0.5$ and mutation probability $q_k = 0.001$ $k = 1, \ldots, 5$. 200 cases and 200 controls were simulated and analyzed. Within columns, LOD and PLOD are calculated for the same data set, and quantiles are estimated with the same random permutations. The horizontal lines show the critical limits for Z_{max} for different significance levels α (displayed on the right y-axes). Marker positions are indicated with dotted, vertical lines, and the true disease location $\tau = 0.36$ with a solid vertical line.

7.2 Power calculations

To estimate the power of the tests we have performed tests for multiple simulated data sets. The results are plotted as a Receiver Operating Characteristic (ROC), i.e. power vs. significance level, see e.g. Bradley (1996). For each of N independent simulations from the genetic model, a p-value $\hat{\alpha}_i$, $i = 1, \ldots, N$, is estimated from the results of Q random permutations as in (3). The power β as a function of α could then be estimated by Monte

Carlo as

$$\hat{\beta}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} 1_{\{\hat{\alpha}_i < \alpha\}}$$

The ROC displayed in Figure 7 is an estimation based on 100 simulated data sets with $K = 10$ markers for 200 cases and 200 controls. The model parameters were $\lambda = 2$, $p = 0.1$ and prevalence 0.001. Each p-value was estimated from $Q = 10000$ permutations, and thus p-values larger than $1{\cdot}10^{-3}$ could be estimated accurately. Simulated datasets also admit calculation of the *golden standard* likelihood. For chromosome i let $M_i = 1$ if $i \in M_s$ and 0 otherwise. By (6) and (7) we get $L_{gold} = \prod_{v=1}^{m} P(M_{2v-1}, M_{2v}|Y_v)$. The p-values are then obtained by the permutation procedure. To cut down computation time for the ROC, while not altering the test performance, $Z(x)$ and PZ(x) were only calculated for $x = 0.2, 0.3, \ldots, 0.6$, i.e. $Z_{\max}$ and $PZ_{\max}$ were based on (P)LOD score at $N_x = 5$ non-marker positions around τ. As the LOD score has a steeper ROC for small α it has better performance than the PLOD score. Despite the relatively weak model, both tests turn out positively in a comparison with the baseline $\beta(\alpha) = \alpha$, corresponding to a test that cannot discriminate between H_0 and H_1. Neither of the tests match the golden standard.

To reach the golden standard, a denser set of markers should be needed. As the computational demand for the LOD score grows fast with the number of included markers K, it is not feasible for, say K larger than 20. As datasets with many markers are believed to be needed to unravel the genetic cause of complex diseases, the PLOD score approximation is suggested. Even for large K, the computation of PLOD score is still affordable, helping to perform tests for data sets with more markers, if the number of included individuals is not too large.

As a last example, Figure 8 displays the ROC for PLOD score from 100 runs with $Q = 10000$ permutations for the same genetic model as in Figure 7 ($p = 0.1$, $\lambda = 2$ and prevalence 0.001). The data set now consists of $K = 25$ markers for 200 cases and 200 controls. $PZ_{\max}$ is calculated from PZ at the same 5 positions in the vicinity of τ, as in Figure 7. The LOD score is not computationally tractable, but to use all 25 markers is possible with the PLOD score approximation. Comparison with Figure 7 shows that the performance of PLOD score has improved, but that it still gives worse results than LOD-score did with $K = 10$ markers. In other simulations with higher disease allele frequency $p = 0.2$ the quality of PLOD calculated from $K = 10$ markers almost matched that of LOD score with $K = 10$. For situations where calculation of LOD-score is not feasible, PLOD-score could be a potentially useful approximation.

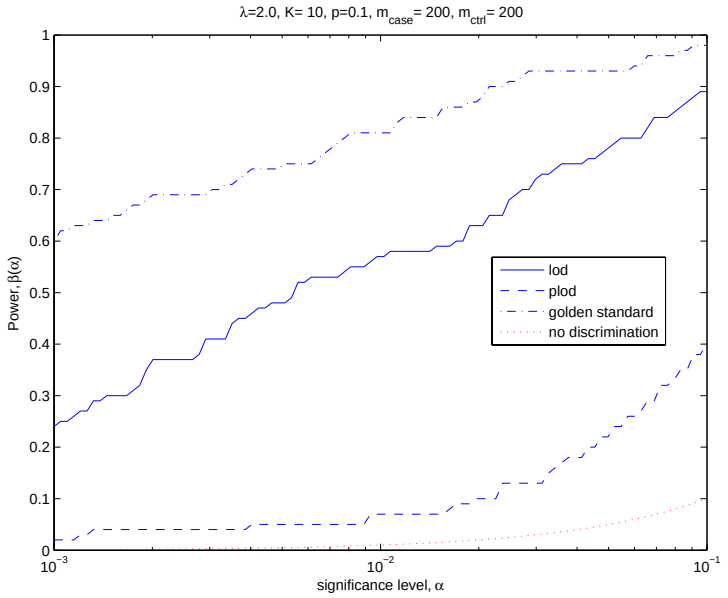

Figure 7: ROC calculated from $N = 100$ p-values, each calculated from $Q = 10000$ permutations. Genetic model is multiplicative with relative risk $\lambda = 2$, disease allele frequency 0.1, $K = 10$, $\rho = 1.5$, marker allele frequency $f_k(1) = 0.5$ and mutation probability $q_k = 0.001$ $k = 1, \ldots, 10$. 200 cases and 200 controls were simulated and analyzed.

7.3 Time consumption

Table 1 contains the mean computation times for the accounted LOD and PLOD scores. Computations were performed on one of the processors of a fast computer, a AMD Athlon(tm) 64 X2 Dual Core Processor 5000+ with 2.6GHz processor and total memory 2GB.
The computation times include permutation testing, and although highly dependent on the computer used, they demonstrate that the tests are feasible even for quite large data sets with many markers. If implementation was done in a program language such as `c++` instead of `Matlab`, we believe that the computation times could be considerably lower. Comparing the empirical results to the theoretical complexity calculations of Section 6.2 and 6.1 give quite large deviations. We believe this is mainly an artefact of memory constraints, which prevent us from proper vectorization of the code for the permutation test of the PLOD score. Also this would be avoided in a programming language that is effective for heavy computations.

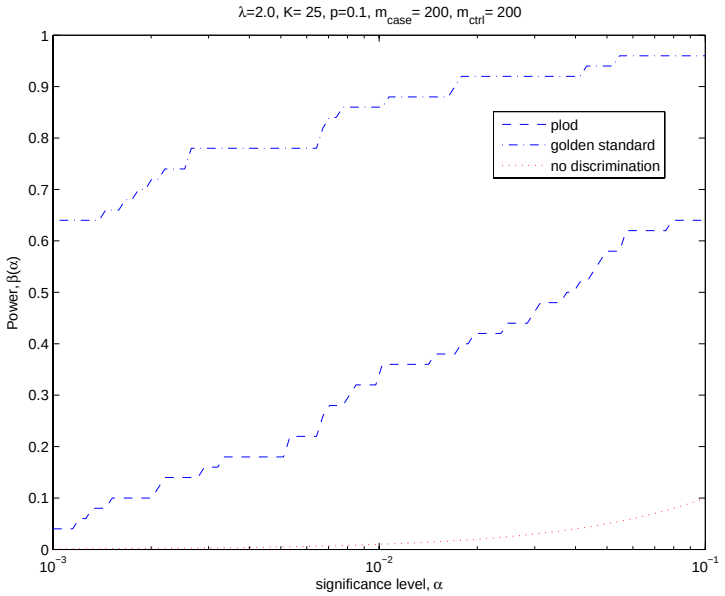

Figure 8: ROC calculated from $N = 100$ p-values, each calculated from $Q = 10000$ permutations. Genetic model is multiplicative with relative risk $\lambda = 2$, disease allele frequency 0.1, $K = 25$, $\rho = 1.5$, marker allele frequency $f_k(1) = 0.5$ and mutation probability $q_k = 0.001$ $k = 1, \ldots, 25$. 200 cases and 200 controls were simulated and analyzed.

8 Discussion

For gene-mapping studies, ARGs have been used as a model framework of chromosome evolvement of a sample, thus assigning probabilities to different scenarios of the relatedness of the sample chromosomes. In this article we extend the ARG, to better describe the genealogy of a retrospective sample of chromosomes. To achieve this, we define two subpopulations, of mutated and unmutated chromosomes respectively, which have different recombination or coalescence rates. Specifically, the rates of coalescence and recombination change in time, depending on the number of mutated/unmutated chromosomes in the population at a certain time. As a model for chromosome samples, the retrospective ARG is general, as it handles arbitrary genetic models, adapts to marker allele frequencies and copes with neutral mutations. The model can allow for varying population size $N(t)$ and disease allele frequency $p(t)$ through the parameters N_M/N, κ and κ_M. Stochastic disease allele frequency $p(t)$ can be handled viewing $p(\cdot)$ as a hidden variable

parameters				time (s)		
m	K	N_x	Q	LOD	PLOD	Figure
400	5	20	100000	150	61000	Figure 6
400	10	5	10000	1000	1600	Figure 7
400	25	5	10000	—	1800	Figure 8

Table 1: Mean computation times for different sample sizes. The mean computation times for LOD and PLOD are measured in seconds.

of the ARG which is repeatedly simulated according to a population genetic model and then, conditionally on $p(\cdot)$ the ARG is simulated as described in Section 4.
The main contribution of the article is that we show how a retrospective ARG can be used to calculate likelihood and LOD score, i.e. we use the retrospective ARG directly for multipoint gene-mapping.

For the purpose of simulation from an ARG (with slight approximations) there are various software available, see e.g. Hudson (2002) and Marjoram and Wall (2006). The simulations include recombinations and neutral mutations, and can be modified to adapt to varying recombination or mutation rates. Thus, there are good methods to generate random samples from a population. However, these algorithms provide no good way to obtain the kind of highly non-random samples that are used for case-control association studies, linkage-disequilibrium mapping studies and other gene-mapping purposes. We therefore believe one important application of our retrospective ARG is simulation of haplotype data $\boldsymbol{h}$ conditional on $\boldsymbol{Y}$. This can be achieved by first simulating M_s conditional on $\boldsymbol{Y}$, then $\mathcal{A}$ conditional on M_s, $\boldsymbol{h}'$ conditional on n' with specified founder haplotype frequencies, and finally $\boldsymbol{h}$ conditional on $\mathcal{A}$ and $\boldsymbol{h}'$. In this way we mimic a non-random sample where chromosomes carrying a disease mutation tend to be more closely related than chromosomes not carrying the disease, and thus also more closely related than the population as a whole. Ascertainment corrected simulation has earlier been proposed by Zöllner and von Haeseler (2000) and recently by Wang and Rannala (2004, 2005), whose models show large similarities with the one we propose. Just as in this paper, the model of Wang and Rannala (2005) handles incomplete penetrance and genotype data. However, they use discrete generations, and account for varying or stochastic disease allele frequency. Varying population size $N(t)$ and disease allele frequency $p(t)$ is incorporated also in our model, and are controlled by parameters κ, κ_M and N_M/N for a model of exponential growth. That the disease locus of Zöllner and von Haeseler (2000) and Wang and Rannala (2004, 2005) is assumed to

be on the same side of all markers, is probably easy to generalize to the case we use, with markers on both sides of the disease locus. Further the SNP locations are simulated as part of the procedure, whereas ours appear at predetermined positions. This simpler approach allows researchers interested in a specified region to choose the marker locations among the SNP locations in the human genome, which are nowadays easily available, e.g. from the HapMap project (The International HapMap Consortium, 2003).

A deficit with our model is that the nuisance parameters ξ, including the recombination rate r, number of generations since the mutation G_M, and since founder generation G, is not part of the estimation procedure, but must be set by the modeller. In many real life situations these parameters are not known, and must in this case be estimated before the analysis is performed. Thus, the most important extension of our method would possibly be to include estimation of the nuisance parameters.

In practice, computational complexity is an important issue for simulation, but even more for likelihood calculation. The reason is that the number of ARGs compatible with sample data is enormous for most study designs. Without model approximations, as in Section 5, exact likelihood computation (5) is not feasible, because of the daunting summation over $\boldsymbol{h}$, $\boldsymbol{h}'$, $\mathcal{A}$ and M_s. The direct Monte Carlo approach would be to generate random replicates of $(\boldsymbol{h}, \boldsymbol{h}', \mathcal{A}, M_s)$ conditional on $\boldsymbol{Y}$ as described in the previous paragraph. However, this is not a feasible approach in general since the integrand $P(\boldsymbol{g}|\boldsymbol{h})$ would be zero with very large probability. A remedy is to use importance sampling, i.e. to sample $(\boldsymbol{h}, \boldsymbol{h}', \mathcal{A}, M_s)$ from another distribution that mimics the conditional distribution given $(\boldsymbol{g}, \boldsymbol{Y})$. Fearnhead and Donnelly (2001, 2002) have considered importance sampling conditional on $\boldsymbol{g}$ in the context of estimating recombination rates r. An interesting topic would be to extend their algorithms to our retrospective context of disease locus estimation

To give an example of a computationally feasible LOD score based on the retrospective ARG we introduce model approximations. The resulting model, based on star topology for cases and independence for controls, is a severe simplification compared to the original. This approximation has earlier been considered for calculation of LOD score by e.g. Terwilliger (1995) and Service et al. (1999). Compared to these however, our resulting LOD score is an important generalization since it handles more markers and unknown phase. Further, for binary phenotypes, we can obtain p-values by a computationally feasible permutation algorithm.
Of the approximations, assumption vi, that mutations occur according to a Markov process, is not essential for the method, but is included for convenience. Further, for the LOD score calculations it is implicitly assumed that

$\tau \neq x_k$, $k = 1, \ldots, K$. Although computations will not collapse if LOD score is calculated at marker loci, there will in general be a discrepancy between disease allele frequency and marker allele frequency that will hamper the results. More precisely, even for calculations at a marker locus, the procedure does not require that exactly the chromosomes with disease mutation should have a certain allele.

As a computationally more tractable alternative to ARG methods, haplotype-clustering methods have been suggested (e.g. Molitor et al. (2003), Durrant et al. (2004) and Waldron et al. (2006)). There a clustering, based on a chosen haplotype similarity measure, form a cladogram which, compared to the ARG, is a coarse approximation of population evolution. The suggested similarity metrics have been simple, quite ad hoc, and based on identity by state (IBS), e.g. the largest shared region between two chromosomes around a putative disease locus, possibly normalized for varying allele frequencies. The retrospective ARG models which regions that the sample chromosomes share IBD. By utilizing this for two chromosomes at a time, the model hold promise for calculating an IBD-based similarity metric, which incorporates disease status, copes with neutral mutations and adapts to allele frequencies. We hope that this can bridge the gap between haplotype similarity methods and ARG-based methods, and plan to explore this in a forthcoming paper.

Appendix

Regularity conditions on exponential growth model imposed by i-iv. For the exponential growth model (4), Conditions i-iv imply (recall that $T_M = 1$)

$$\begin{aligned} p(t) &= (N_M/N)\exp(G_M(\kappa - \kappa_M)t) \\ &= (N_M/N)^{T_M - t}(1/N(T_M))^t, \\ \int_0^{T_M} \mu_U(t)dt &\approx (\exp(G_M \kappa T_M) - 1)/(N\kappa) \\ &\approx 1/(N(T_M)\kappa), \\ \mu_M(t) &= G_M \log(1 - \exp(\kappa_M G_M(t-1)))^{-1} \\ &\leq G_M \log(1 - \exp(-\kappa_M G_M \varepsilon))^{-1}, \end{aligned}$$

where the two approximations of the middle approximation requires $N_M \gg 1$ and $N_M(t) \ll N(t)$ at all time points $0 \leq t \leq T_M$, and the last inequality applies for all $0 \leq t \leq T_M$. Hence for conditions i-iv to hold it suffices that

$$\begin{aligned} N_M &\ll N, \\ N(T_M) &\gg \max(\kappa^{-1}, 1), \\ G_M \exp(-\varepsilon \kappa_M G_M) &\ll 1. \end{aligned}$$

In words, the disease allele frequency $p = N_M/N$ should be small, the founder population large, and either G_M or κ_M large. □

HMM algorithm for computing likelihood ratio (20) and (21). Consider a subset I of $\{1, \ldots, n\}$ and let $R = D \cap (I \times \{1, \ldots, K\})$ be the set of mutated markers (i, k) for chromosomes $i \in I$ and $C_k = C_{x_k}$ the k^{th} column of R, $k = 1, \ldots, K$. Under the assumption that $\tau = x$, we will devise a HMM algorithm for computing

$$S = E_x \left(\prod_{k=1}^{K} U_k(C_k) | \boldsymbol{Y}_V \right), \tag{A.1}$$

where $\boldsymbol{Y}_V = \{Y_v,\, v \in V\}$, $V = \{v; 1 \le v \le m, I \cap \{2v-1, 2v\} \neq \emptyset\}$ and $U_k(C_k) = U_k(C_k; \boldsymbol{g})$ a given function. We will apply this when i) S equals the v^{th} term of the outer product in (20) (with $I = \{2v-1, 2v\}$, $V = \{v\}$ and U_k the k^{th} term of (18), with R_v in place of D) and ii) when $S = \text{LR}(x)$ is the likelihood ratio, expanded as in (21) (with $I = \{1, \ldots, n\}$, $V = \{1, \ldots, m\}$, $\boldsymbol{Y}_V = \boldsymbol{Y}$ and U_k the k^{th} term of (22)). With obvious modifications of $\boldsymbol{Y}$ and $\boldsymbol{g}$, (A.1) also applies to calculating the relevant terms of the pseudo likelihood ratio $\text{PLR}(x)$.

To begin with, we establish a Markov property of $\{C_k\}$. Because of (12), the columns of R to the left and right of the (assumed) disease locus, $\{C_k\}_{k=k_0}^{1}$ and $\{C_k\}_{k=k_0+1}^{K}$, evolve as two Markov chains with state space all subsets of I. The two chains are independent conditional on their starting values, which depend on $C_x := M_s \cap I$. As the chains progress, C_k is non-increasing in both directions, with Lineage $i \in I$ lost at x_k when there are recombination events at X_i^- (X_i^+) just to the right (left) of x_k affecting i.

Since recombinations occur as independent Poisson processes with rate $\rho\pi(\cdot)$ along different mutated i-lineages from $t = 0$ to $t = T_M = 1$, it is easy to see that $\{X_i^-\}_{i \in M_s}$ and $\{X_+^-\}_{i \in M_s}$ are independent random variables with

$$\begin{array}{rcl} P_x(X_i^- < x' | i \in M_s) & = & \exp\left(-\rho \int_{x'}^{x} \pi(y)dy\right),\ 0 \le x' < x. \\ P_x(X_i^+ > x' | i \in M_s) & = & \exp\left(-\rho \int_{x}^{x'} \pi(y)dy\right),\ x < x' \le 1. \end{array} \tag{A.2}$$

Let F^- and F^+ denote the distribution functions of X_i^- and X_i^+, and put

$$\begin{array}{rcll} r_k & = & (F^-(x_k) - F^-(x_{k-1}))/F^-(x_k), & 1 \le k \le k_0, \\ r_k & = & (F^+(x_{k+1}) - F^+(x_k))/(1 - F^+(x_k)), & k_0 + 1 \le k \le K, \\ r_x^- & = & F^-(x) - F^-(x_{k_0}), & \\ r_x^+ & = & F^+(x_{k_0+1}) - F^+(x). & \end{array}$$

In words, when $k \le k_0$, r_k is the probability of a recombination between x_{k-1} and x_k given that no recombination has occurred between x_k and x. r_x^- is

the probability of a recombination event between x_{k_0} and x for a mutated lineage $i \in M_s$. The interpretation of r_k for $k \geq k_0 + 1$ and r_x^+ is similar. This gives rise to transition probabilities

$$\begin{aligned}
P(C_{k-1} = C'|C_k = C) &= r_k^{|C|-|C'|}(1-r_k)^{|C'|},\ k = 1, \ldots, k_0, \\
P(C_{k+1} = C'|C_k = C) &= r_k^{|C|-|C'|}(1-r_k)^{|C'|},\ k = k_0+1, \ldots, K, \\
P_x(C_{k_0} = C'|C_x = C) &= (r_x^-)^{|C|-|C'|}(1-r_x^-)^{|C'|}, \\
P_x(C_{k_0+1} = C'|C_x = C) &= (r_x^+)^{|C|-|C'|}(1-r_x^+)^{|C'|},
\end{aligned}$$

provided $C' \subseteq C$ (otherwise the transition probabilities are zero). Define

$$S_k(C) = \begin{cases} E_x(\prod_{l=1}^k U_l(C_l)|C_k = C), & 0 \leq k \leq k_0, \\ E_x(\prod_{l=k}^K U_l(C_l)|C_k = C), & k_0 + 1 \leq k \leq K+1. \end{cases}$$

and

$$S(C) = E_x(\prod_{l=1}^K U_l(C_l)|C_x = C).$$

Then, the recursive algorithm for computing S can be formulated in terms of $S_k(C)$, $S(C)$ and the transition probabilities as follows: Start with initial conditions

$$\begin{aligned}
S_0(C) &= 1, \\
S_{K+1}(C) &= 1,
\end{aligned} \quad \forall C \subseteq I.$$

Then, define recursively for all $C \subseteq I$

$$\begin{aligned}
S_k(C) &= \textstyle\sum_{C' \subseteq C} S_{k-1}(C')U_k(C)P(C_{k-1} = C'|C_k = C),\ k = 1, \ldots, k_0, \\
S_k(C) &= \textstyle\sum_{C' \subseteq C} S_{k+1}(C')U_k(C)P(C_{k+1} = C'|C_k = C),\ k = K, \ldots, k_0+1.
\end{aligned}$$

The final two steps are

$$\begin{aligned}
S(C) &= \left(\textstyle\sum_{C' \subseteq C} S_{k_0}(C')P(C_{k_0} = C'|C_x = C)\right) \\
&\quad \cdot \left(\textstyle\sum_{C' \subseteq C} S_{k_0+1}(C')P(C_{k_0+1} = C'|C_x = C)\right),\ \forall C \subseteq I \\
S &= \textstyle\sum_{C \subseteq I} S(C)P(C_x = C|\boldsymbol{Y}_V).
\end{aligned}$$

In the last step we use (7) to evaluate $P(C_x = C|\boldsymbol{Y}_V)$. The total complexity of the algorithm is $O(K2^{|I|})$, where $2^{|I|}$ is the state space size. □

References

Bradley, A. (1996). Roc curves and the χ^2 test. *Pattern Recognition Letters*, 17(3):287–294.

Durrant, C., Zondervan, K. T., Cardon, L. R., Hunt, S., Deloukas, P., and Morris, A. P. (2004). Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am J Hum Genet*, 75(1):35–43.

Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics*, 159(3):1299–1318.

Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64:657–680.

Griffiths, R. and Marjoram, P. (1997). An ancestral recombination graph. In Donnelly, P. and Tavaré, S., editors, *Progress in Population Genetics and Human Evolution*, pages 257–270. Springer, New York.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, 23(2):183–201.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338.

Kingman, J. (1982a). The coalescent. *Stoch Proc Appl*, 13:235–248.

Kingman, J. (1982b). On the genealogy of large populations. In Gani, J. and Hannan, E., editors, *Essays in Statistical Science: Papers in Honours of P.A.P. Moran*, Journal of applied probability, pages 97–112. North-Holland Publishing, Amsterdam.

Kraft, P. and Thomas, D. (2000). Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet*, 66(3):1119–31.

Larribe, F., Lessard, S., and Schork, N. J. (2002). Gene mapping via the ancestral recombination graph. *Theor Popul Biol*, 62(2):215–229.

Marjoram, P. and Wall, J. D. (2006). Fast "coalescent" simulation. *BMC Genet*, 7:16.

The International HapMap Consortium (2003). The International HapMap Project. *Nature*, 426(6968):789–796.

McPeek, M. and Strahs, A. (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet*, 65(3):858–75.

Molitor, J., Marjoram, P., and Thomas, D. (2003). Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am J Hum Genet*, 73(6):1368–84.

Service, S. K., Lang, D. W., Freimer, N. B., and Sandkuijl, L. A. (1999). Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am J Hum Genet*, 64(6):1728–1738.

Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am J Hum Genet*, 56(3):777–787.

Thomas, D. C., Stram, D. O., Conti, D., Molitor, J., and Marjoram, P. (2003). Bayesian spatial modeling of haplotype associations. *Hum Hered*, 56(1-3):32–40.

Waldron, E. R. B., Whittaker, J. C., and Balding, D. J. (2006). Fine mapping of disease genes via haplotype clustering. *Genet Epidemiol*, 30(2):170–179.

Wang, Y. and Rannala, B. (2004). Simulating a coalescent process with recombination and ascertainment. In Istrail, S., Waterman, M., and Clark, A., editors, *Computational Methods for SNPs and Haplotype Inference*, volume 2983 of *Lecture Notes in Bioinformatics*, pages 84–95. Springer.

Wang, Y. and Rannala, B. (2005). In silico analysis of disease-association mapping strategies using the coalescent process and incorporating ascertainment and selection. *Am J Hum Genet*, 76(6):1066–1073.

Wiuf, C. and Hein, J. (1999). The ancestry of a sample of sequences subject to recombination. *Genetics*, 151(3):1217–1228.

Zöllner, S. and von Haeseler, A. (2000). A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. *Am J Hum Genet*, 66(2):615–628.